Speech intelligibility in cars: the effect of speaking style, noise and listener age

Speech intelligibility in cars: the effect of speaking style, noise and listener age

Cassia Valentini Botinhao[1], *Junichi Yamagishi*[1,2,3]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK
[2] National Institute of Informatics, Japan
[3] SOKENDAI University, Japan

cvbotinh@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

Intelligibility of speech in noise becomes lower as the listeners age increases, even when no apparent hearing impairment is present. The losses are, however, different depending on the nature of the noise and the characteristics of the voice. In this paper we investigate the effect that age, noise type and speaking style have on the intelligibility of speech reproduced by car loudspeakers. Using a binaural mannequin we recorded a variety of voices and speaking styles played from the audio system of a car while driving in different conditions. We used this material to create a listening test where participants were asked to transcribe what they could hear and recruited groups of young and older adults to take part in it. We found that intelligibility scores of older participants were lower for the competing speaker and background music conditions. Results also indicate that clear and Lombard speech was more intelligible than plain speech for both age groups. A mixed effect model revealed that the largest effect was the noise condition, followed by sentence type, speaking style, voice, age group and pure tone average.
Index Terms: Speech intelligibility, age, hearing loss, car noise

1. Introduction

The use of speech interfaces in cars over visual displays increases safety as drivers are not required to look away from the road [1]. This is particularly relevant for older adults as this group of drivers tend to focus more on the road, spending less time looking at in-vehicle displays [2]. Thus, using speech as a way to provide information to older drivers seems like a good choice. However, older drivers are more likely to experience age-related and noise-induced elevations of auditory thresholds, as well as increased mental workload.

Elderly listeners experience difficulties processing speech, particularly in noise and under stress. This holds true even when there is no evidence of abnormal hearing thresholds [3]. It has been argued [4] that this could be due to temporal resolution loss caused by disrupted neural connections that happen with age. According to [4], this kind of loss appears as a decline in the processing of the slowly varying envelope (as observed by measuring performance on gap detection tasks [5]) and the processing of the fine structure envelope (as observed by measuring performance on frequency modulation detection [6], pitch discrimination [7], inter-aural phase and time difference detection tasks [8]). This impacts speech understanding in noise, but particularly in fluctuating noises, as good temporal acuity is necessary for a listener to take advantage of the relatively silent gaps in the noise. In [5] it was found that word recognition in competing babble correlated significantly with temporal resolution and age, but not with absolute hearing loss. In [9], however, it was found that in stationary noise, hearing loss significantly contributed to explaining differences in speech reception and no other predictors (age, temporal acuity) seemed to contribute.

It is possible to modify speech in such a way that the mixture of speech and noise is more intelligible for the listener without an overall level increase. One could for instance modify speech produced in quiet conditions by promoting acoustic changes observed in speaking styles that are more intelligible in noise conditions, such as clear speech (produced with the intent to counter adverse listening conditions) and Lombard speech (produced in noise). A study [10] that investigated the driving performance of a group of university students found that navigation systems with dominant voices (faster speech rate, higher amplitude and pitch, more pitch variation and dominant messages) lead drivers to follow instructions better. Both clear and Lombard speech have been shown to increase intelligibility for older adults [11, 12, 13, 14], although some studies did not find the clear speech benefit for hearing impaired older adults [15].

In this paper, we are interested in finding which driving conditions and which speaking styles are more intelligible for older adults. For this purpose we have recorded a database of a variety of voices and speaking styles in many different driving conditions. We then performed a listening experiment on a selected portion of the data to gather intelligibility scores from young (below 30 years of age) and old (above 50) adults.

2. Database

2.1. Studio recordings

We recorded two native English speakers, a man and a woman, recruited via an advert in the University of Edinburgh. The selected participants were not professional speakers but were selected according the quality of their recordings from a pool of eight applications. The recordings took place in semi-anechoic chambers. The recording sentence material was: 200 sentences from newspapers selected for the purpose of training text-to-speech voices [16], the first 50 sentences from the Harvard sentences corpus [17] and 50 sentences consisting of instructions on driving a particular route in Edinburgh, UK. The navigation style sentences were mostly made of a verb in imperative form, followed by an adverb or a preposition and a noun. For example: “Turn right onto the Royal Mile”.

Each person was asked to read all 300 sentences in four different speaking styles, which we refer to as: plain, confident, clear and Lombard. The plain style was recorded by asking participants to speak in their normal reading voice. The confident style was created by asking participants to talk as if giving instructions to someone. For the clear style we asked participants to speak as if they were talking to a hearing impaired person. To create Lombard speech, participants were asked to wear headphones. Through these headphones we played a continuous noise signal and participants were asked to read the material as if trying to communicate through that noise. The noise selected

was the car noise recordings from the Demand database [18].

2.2. In car recordings

To record the in-car database we used the B&K 4100 head and torso mannequin. The mannequin was placed in the front passenger seat, fixed with seat belts and the B&K WA-1647 car seat fixture. The recordings took place in Tokyo, Japan, across several days, following one of the two routes: one following city roads and one following a highway. We also recorded GPS location and speed information as well as video from a camera pointing towards the front. For all sessions we used the same hybrid (electric/petrol) compact car, a Toyota Aqua.

2.2.1. Speech recordings

Speech material recorded inside the car included not only the two speakers in four different speaking styles but also data from a professional voice talent as a reference of high quality speaking style data. This extra data consisted of speech from a male British speaker in two speaking styles: plain and Lombard, as described in [19]. In total there were ten voice styles belonging to the three different speakers. The Harvard and the navigation style sentences spoken by each voice was burned to a CD with ten tracks of 100 sentences each, totaling 1000 sentences. The sentences were concatenated with a one second silence in between, and before each track we added a short 1 kHz tone to guide sentence segmentation later on. The CD was played using the car audio system and loudspeakers to simulate satnav generated speech in the car. There were four loudspeakers in total: two in the front and two in the back, located on the doors.

The material was recorded in 10 different sessions that varied in terms of the type of route taken, the weather condition, the background noise in the car and whether any windows were open or not. The driving conditions in the city route were: windows open (WO), windows closed (WC), windows closed with rain (RAIN) and windows closed with a competing speaker (CS). The highway conditions were: WC, RAIN, CS and windows closed with background music (BGM). The parking conditions were: WC and windows closed with hazard lights on (HZ). In all conditions except WO, the air conditioner of the car was on and always at the same level. The competing speaker condition was created by playing pre-recorded speech material from a different voice using a loudspeaker positioned in the backseat of the car. The loudspeaker was positioned at a particular height so as to simulate a person sitting in the middle of the backseat. The competing speaker was always the opposite gender to the voice playing in the car loudspeakers. The background music condition was created by using a CD with tracks whose left channel contained the speech material and whose right channel contained music, to simulate someone listening to music while also listening to speech instructions. In the windows open condition, the window closest to the driver was open half way. The parking condition was recorded while parked off a quiet street. A total of 39 hours of data was recorded.

The volume of the CD for the city routes and parking conditions was fixed to a slightly lower level than for the highway routes. The volume levels were chosen according to what we considered to be a reasonable hearing level for normal hearing adults in each route.

2.2.2. Noise only and impulse response recordings

In order to understand the noise profile and estimate the signal-to-noise ratio of the speech recordings we also recorded noise only segments as well as impulse responses inside the car. In a separate recording session, with the car audio system turned off, we recorded ten minutes of audio in each driving condition described previously, except RAIN.

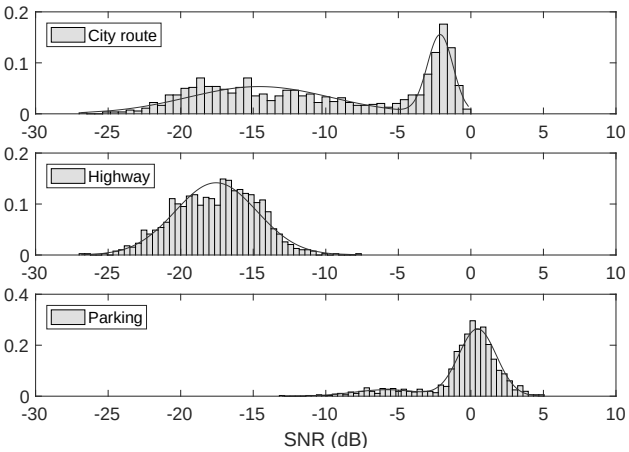

Figure 1: *SNR distributions in the city (top), the highway (middle) and whiled parked (bottom). The continuous line refers to the Gaussian mixture model distribution fitted on the data.*

To estimate the impulse response inside the car we created a sine sweep signal using the tool FuzzMeasure [20]. We recorded different types of responses by playing the sine sweep stimuli first through all of the car loudspeakers and then either only through the left or the right side loudspeakers. We also recorded the sweep signal played over the loudspeaker that was used to simulate the competing speaker. To record all these materials, we parked the car inside an indoor garage with windows closed and air conditioner off in order to minimize the amount of noise as much as possible and improve the impulse response estimation.

2.3. Post processing

2.3.1. Segmentation

After car recordings were done we segmented the data sentence by sentence semi-automatically by cross-correlating the envelope of the clean signal and the envelope of the recorded noisy signal. We noticed that during recording some segments of the data were not recorded due to buffering issues. For this reason 54 sentences were corrupted and could not be segmented properly. The sine sweep recordings were also segmented following a similar procedure. The noise only recordings were segmented manually according to time stamp noted during recording. All material was downsampled to 44.1 kHz and high pass filtered to attenuate noise found below 70Hz.

2.3.2. Impulse response estimation

The car impulse response was estimated from the segmented sine sweep response using FuzzMeasure [20]. The final impulse response was the minimum phase version of the impulse response generated by the tool.

2.3.3. SNR estimation

We estimated the signal to noise ratio (SNR) of each recorded sentence as follows:

$$\text{SNR(dB)} = 20\log_{10}\frac{\alpha E_s}{E_m - \alpha E_s} \quad (1)$$

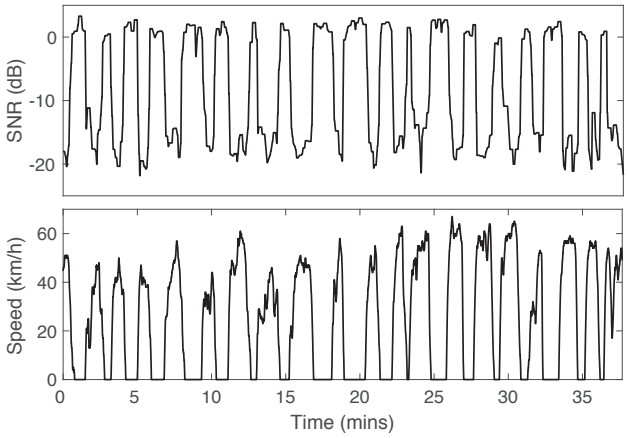

Figure 2: *Sentence level SNR (top) and car speed (bottom) trajectories along the city route.*

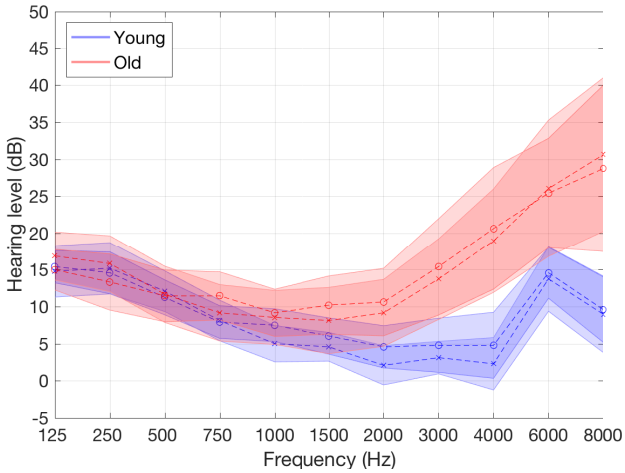

Figure 3: *Average hearing curves of young (blue) and old (red) participants (o: right ear, x: left ear). Shaded area shows one standard deviation.*

where E_s is the root mean square energy (RMS) of the clean speech signal of that particular sentence convolved with the estimated car impulse response, E_m is the RMS of the recorded sentence and α a positive constant value. The α value was chosen so that $E_m - \alpha E_s > 0$ holds true for every sentence.

Fig. 1 shows the SNR distribution of the sentences recorded in the city route, the highway and while parked, in the condition where all windows were closed. We can see that the range of SNR values in the highway route is narrower than the city route, reflecting the fact that there is less variation in noise levels when speed is steadier. In the parking and city route conditions, the SNR distribution follows a bimodal distribution. The two modes in the parking condition reflect the difference in noise levels when the engine switches between electric and petrol. The two modes in the city route distribution reflect the fact that some sentences were recorded while the car was moving and others while the car was stopped either at traffic lights or in traffic jams. Fig. 2 shows the evolution of the sentence level SNR and the speed of the car during a portion of the city route recording session, where we can see that the noise level fluctuation follows a similar pattern to the speed curve: the higher the speed the lower the SNR.

2.3.4. Data selection

As we observed that the SNR levels varied a lot even in the same condition, we decided to exclude sentences where the noise level was either too high or too low. In order to do so we fitted a Gaussian mixture model to the estimated SNR distribution. For the highway conditions only one mixture was used while for the city route and the parking conditions, two mixtures were fitted, as illustrated by the continuous lines in Fig. 1. Sentences whose estimated SNR values were further than one standard deviation away from the mean of a particular Gaussian were excluded. For the city route the chosen Gaussian, whose mean we refer to, was the one with the lowest mean SNR value (highest noise level), which we expect covers the sentences recorded while the car was moving. For the parking conditions the Gaussian chosen was the one with the highest SNR value (lowest noise level), where sentences were recorded while the electric engine was on. The mean SNR values chosen for the city route, highway and parking WC condition displayed in Fig. 1 were -14.6dB, -17.5dB and 0.5dB respectively. Around 44% of the sentences were excluded.

3. Listening test

3.1. Design

The listening test was organized in four blocks of 25 sentences each, a total of 100 different sentences. In each block, only stimuli belonging to a particular voice were used, so that every listener heard examples of four different voices and for every three consecutive listeners all voices were covered. The first five sentences of each block were used to train listeners to get used to that particular voice. The remaining 20 sentences in the block covered all 10 driving conditions, so that two examples of each condition were presented. The order of the sentences was randomized per participant as was the order of the conditions inside each block. Participants were asked to type in what they could hear from each sentence. They could only play each sentence once. Participants were wearing headphones for this task and inputted their responses using a normal keyboard in a Matlab interface.

3.2. Participants

Two groups of English native participants were recruited: 24 younger subjects (mean age: 22.7 years; range: 19 - 29); and 24 older subjects (mean age: 61.1 years; range: 52 - 76). The participants were recruited in Edinburgh and the experiment took place in Edinburgh as well. All participants reported not being aware of any severe hearing problems. To assess this, prior to the listening experiment, we performed a hearing test using an audiometer, following the procedure described in [21]. The average hearing curves of each group are shown in Fig. 3. We can see that the hearing level curves of the age groups differ most at high frequencies, above 3 kHz.

3.3. Results

We present the results in terms of word accuracy, calculated as the percentage of correctly identified words on a per sentence basis, excluding common words and following the procedure described in [19]. The participants' responses were checked for typos and misspellings before we computed word accuracy values. To test for significance, we used a Mann-Whitney U test at a p-value of 0.05 with a Holm Bonferroni correction due to the large number of pairs to compare.

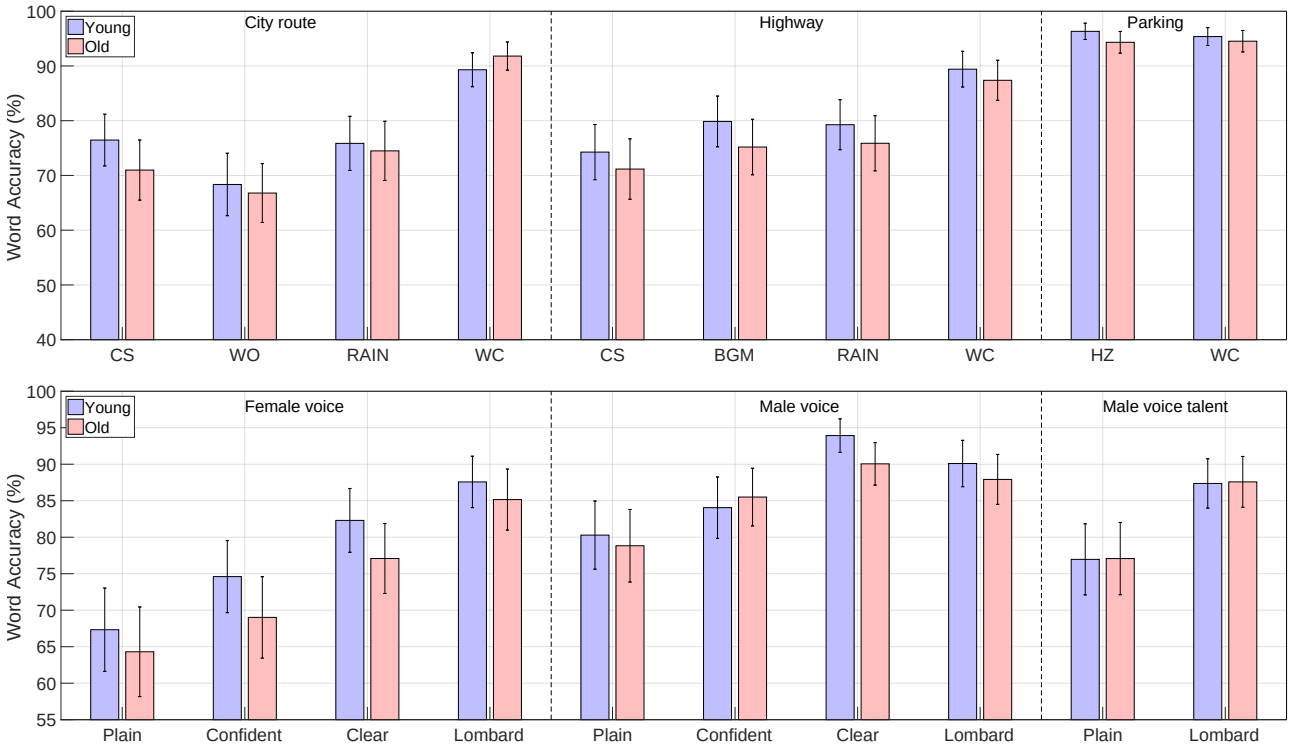

Figure 4: *Average word accuracy rates and confidence intervals for each driving condition (top) and voice type (bottom).*

To analyse the underlying size effects we model the word accuracy data using a linear mixed effects model [22]. The following random factors were considered: age group (two levels), pure tone average (PTA) (continuous value), driving condition (ten levels), speaker (three levels), speaking style (four levels) and sentence type (two levels). PTA was calculated as the average hearing level in the frequencies of 0.5k, 1k, 2k and 4k Hz. Although other criteria exist, according to the World Health Organization, a PTA value larger than 25 dB is an indicator of hearing loss [23]. Sentence type refers to whether a sentence is a Harvard sentence or navigation style sentence.

We found that the size of all effects were significantly greater than zero at a 95% confidence interval, i.e. all effects influenced intelligibility. The effect sizes were, however, not significantly different from each other. The largest effect found was the driving condition, followed by sentence type, speaker, speaking style, PTA and age group. We believe that the sentence type effect was large because the navigation style sentences were relatively more intelligible than the Harvard sentences, possibly due to their highly predictable structure and due to the presence of common Edinburgh road names.

Fig. 4 shows the average word accuracy and standard error in each driving condition, at the top, and for each voice (speaker and speaking style), at the bottom. We can see that the least intelligible driving condition was in the city route with windows open and the easiest conditions, as expected, were both parking conditions and the conditions where windows were closed and no competing speaker or rain noise was present. The performance of young and old participants differed most in the competing speaker and background music conditions, where the accuracy of old participants is smaller, although not significant. This result is inline with research that found fluctuating noise is particularly hard for older adults with normal hearing [5].

In terms of the voices, see Fig. 4 bottom plot, we can see that for both age groups the plain style was the least intelligible, resulting in the lower values of word accuracy. The intelligibility benefit of the other speaking styles was seen for all speakers (even for the non-professional speakers) and for listeners of the two age groups. The higher values of word accuracy were obtained with the Lombard and clear speech styles, which were found to be significantly more intelligible than plain speech for both age groups, apart from the female speaker's clear speech. Lombard speech was more intelligible than clear speech for the female speaker, but the opposite was found for the male speaker. Finally, we can see that the female voice was less intelligible in all equivalent cases, though it also displayed the greatest improvements across speaking styles. We observed that in the easier conditions, the WC cases, the female plain voice was as intelligible as both male plain voices, which indicates that the overall drop in performance seen in Fig. 4 was caused by the other conditions where an additional noise source was present.

4. Conclusions

We were interested in finding which driving conditions and speaking styles influenced speech intelligibility the most for older adults. For this purpose we recorded speech reproduced by car loudspeakers in a range of driving conditions, such as in the city, on the highway, with a competing speaker and with background music. We used this material to construct a listening test where we collected transcription word accuracy rates from both young and older adults. We found that older participants intelligibility scores were lower for competing speaker and background music. Results also indicate that clear and Lombard speech were significantly more intelligible than plain speech for both age groups.

Acknowledgements: The work presented in this paper was partially supported by the TOYOTA motor cooperation.

5. References

[1] A. Barón and P. Green, "Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review," University of Michigan, Transportation Research Institute, Tech. Rep., 2006.

[2] I.-M. Jonsson, M. Zajicek, H. Harris, and C. Nass, "Thank you, I did not see that: in-car speech based information systems for older adults," in *Proc. CHI*. ACM, 2005, pp. 1953–1956.

[3] K. S. Helfer and R. L. Freyman, "Aging and speech-on-speech masking," *Ear and Hearing*, vol. 29, no. 1, p. 87, 2008.

[4] T. Schoof and S. Rosen, "The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners," *Frontiers in Aging Neurosc.*, vol. 6, p. 307, 2014.

[5] K. B. Snell, "Age-related changes in temporal gap detection," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2214–2220, 1997.

[6] N.-J. He, J. H. Mills, and J. R. Dubno, "Frequency modulation detection: effects of age, psychophysical method, and modulation waveform," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 467–477, 2007.

[7] C. Füllgrabe, "Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss," *Am. J. of Audiology*, vol. 22, no. 2, pp. 313–315, 2013.

[8] J. H. Grose and S. K. Mamo, "Processing of temporal fine structure as a function of age," *Ear and Hearing*, vol. 31, no. 6, p. 755, 2010.

[9] E. L. George, A. A. Zekveld, S. E. Kramer, S. T. Goverts, J. M. Festen, and T. Houtgast, "Auditory and nonauditory factors affecting speech reception in noise by older listeners," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2362–2375, 2007.

[10] M. Jonsson and N. Dahlbäck, "In-car information systems: Matching and mismatching personality of driver with personality of car voice," in *Proc. ICHCI*, 2013, pp. 586–595.

[11] D. J. Schum, "Intelligibility of clear and conversational speech of young and elderly talkers," *J. Am. Acad. of Audiology*, vol. 7, pp. 212–218, 1996.

[12] K. S. Helfer, "Auditory and auditory-visual recognition of clear and conversational speech by older adults," *J. Am. Acad. of Audiology*, vol. 9, pp. 234–242, 1998.

[13] M. S. Sommers, N. Tye-Murray, and B. Spehar, "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults," *Ear and Hearing*, vol. 26, no. 3, pp. 263–275, 2005.

[14] M. Fitzpatrick, J. Kim, and C. Davis, "Auditory and auditory-visual lombard speech perception by younger and older adults." in *Proc. AVSP*, 2013, pp. 105–110.

[15] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 259–271, 2002.

[16] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA*, Nov 2013.

[17] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[18] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[19] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, pp. 572–585, 2013.

[20] SuperMegaUltraGroovy, "Fuzzmeasure." [Online]. Available: http://supermegaultragroovy.com/products/fuzzmeasure/

[21] "Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking," British Society of Audiology, Recommended Procedures, December 2015.

[22] T. A. Snijders and R. Bosker, "Multilevel analysis." 2011.

[23] M. C. Colin Mathers, Andrew Smith, "Global burden of hearing loss in the year of 2000," World Health Organization, Report, 2000.